5Н

УДК 004.8

Н.А. Новоселова, И.Э. Том
Объединенный институт проблем информатики
Национальной академии наук Беларуси, г. Минск
novosel@newman.bas-net.by, tom@newman.bas-net.by

Метод оценки кластерной структуры и кластеризации данных

В статье рассматривается проблема разработки методов кластеризации, которые являются устойчивыми к инициализации (количество кластеров и начальные параметры кластеров), к различным по объему кластерам, к выбросам в данных. Предлагается метод оценки кластерной структуры и кластеризации данных, который основан на расчете значений близости объектов данных в многомерном признаковом пространстве. Метод является устойчивым к инициализации параметров кластеризации, к выбросам в данных и позволяет определять кластерную структуру и количество кластеров в ходе самоорганизации объектов данных.

Введение

Одним из первых шагов при анализе набора данных с целью выявления новых знаний или скрытых закономерностей в данных является кластеризация. Процесс кластеризации направлен на определение групп данных, которые являются схожими согласно некоторой мере близости. Последующий анализ кластерной структуры данных позволяет выявить общие функциональные или иные свойства объектов данных, принадлежащие отдельным кластерам, и впоследствии использовать результаты для сокращения размерности, выявления информативных признаков, характеризующих отдельные объекты данных и предсказания свойств новых объектов данных по набору признаков.

Все существующие кластерные методы можно классифицировать на иерархические методы кластеризации [1-3], методы вероятностной кластеризации [4], [5], методы кластеризации, основанные на оптимизации различных целевых функционалов [2], [6], нейросетевые методы кластеризации, как например, самоорганизующиеся карты Кохонена [7], [8], методы кластеризации с использованием теории графов [9], методы, основанные на плотности распределения данных [10], [11] и т.д. Различные методы, разработанные в рамках каждой из вышеперечисленных групп, не являются одинаково эффективными при анализе произвольного набора данных. Каждый из методов имеет преимущество при решении определенного круга практических задач. Наиболее широко на практике используются оптимизационные методы, одним из первых представителей которых является метод четкой кластеризации K-средних (HCM) [12], метод нечеткой кластеризации C-средних (FCM) [6], [13], метод вероятностной кластеризации PCM [14] и метод NC [15]. Однако большинство из разработанных на сегодня кластерных методов в той или иной степени являются неустойчивыми. Устойчивость метода кластеризации включает следующие аспекты [16]: 1) устойчивость к инициализации (количество кластеров и начальные параметры кластеров), 2) устойчивость к различным по объему кластерам (возможность определить кластеры, имеющие разные объемы), 3) устойчивость к выбросам в данных (не должны оказывать влияние на результат кластеризации).

Таким образом, актуальным является разработка таких методов кластеризации, которые являются устойчивым согласно всем трем указанным аспектам и в то же время являются достаточно простыми в вычислительном плане, что позволит использовать их для анализа больших объемов многомерных данных.

Цель работы авторами предлагается **метод оценки кластерной структуры и кластеризации данных (МКК), который основан на оценке близости объектов данных в многомерном признаковом пространстве.** Данный метод является устойчивым к инициализации параметров кластеризации: начальному положению центров кластеров и количеству кластеров, а также к случайным выбросам в данных. Согласно предложенному методу кластерная структура и количество кластеров в данных определяются автоматически в ходе самоорганизации объектов данных. Для определения локально-оптимального количества и состава кластеров используется метод агломеративной иерархической кластеризации [3].

Определение кластерной структуры данных

Основным этапом реализации предложенного МКК является осуществление кластеризации данных путем оптимизации (максимизации) следующего функционала:

$$F(v)=\sum_{i=1}^{c}\sum_{j=1}^{n}\left(\exp-\frac{\left\|x_j-v_i\right\|^2}{\delta}\right)^{\gamma}, \qquad (1)$$

где x_j – j-й объект данных, представляющий собой точку $x_j=\left(x_j^1,x_j^2,\ldots,x_j^m\right)$ в m-мерном признаковом пространстве; $v_i=\left(v_i^1,v_i^2,\ldots,v_i^m\right)$ – центр i-го кластера; c – количество кластеров; δ – дисперсия всего набора данных $X=\left\{x_1,x_2,\ldots,x_n\right\}$; γ – оцениваемый параметр кластеризации.

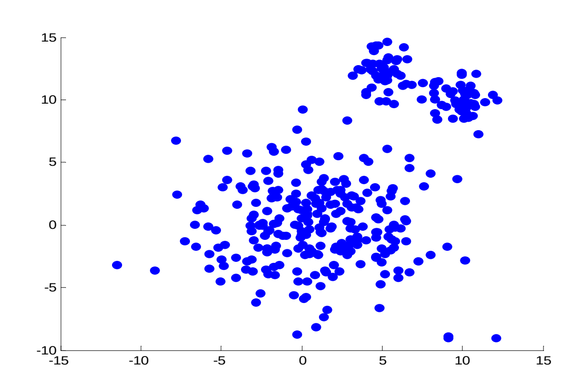


Рисунок 1 – Набор данных Data1, состоящий из трех кластеров

Функция $f(v_i)=\exp-\frac{\left\|x_j-v_i\right\|^2}{\delta}$ представляет собой меру сходства объекта x_j и i-го кластерного центра v_i, которая используется в работе [17]. Таким образом, целью оптимизационного кластерного алгоритма, определяемого функционалом (1), является поиск таких значений центров кластеров $v_i, i=1,\ldots,c$, которые максимизируют полную меру сходства объектов данных и кластерных центров. В настоящей работе исполь-

зуется евклидово расстояние $\left\|x_j - v_i\right\|^2$ между объектом данных x_j и кластерным центром v_i, однако вместо него можно использовать любую приемлемую меру расстояния.

Основное внимание необходимо обратить на выбор параметра γ функционала (1). Данный параметр позволяет определить положение локальных экстремумов функционала (1) и таким образом оценить плотность распределения данных в окрестности каждого объекта набора данных. Для иллюстрации влияния значения параметра γ на положение локальных экстремумов функционала (1) рассмотрим набор данных Data1, представленный на рис. 1.

Рассчитаем для каждого объекта $x_k, k = 1,\ldots,n$ набора данных Data1 значение следующей функции, определяющей общую меру сходства объекта x_k со всеми остальными объектами данных:

$$F(x_k) = \sum_{j=1}^{n}\left(\exp -\frac{\left\|x_j - x_k\right\|^2}{\delta}\right)^{\gamma}. \qquad (2)$$

Значение функции $F(x_k)$ в точке x_k можно считать оценкой плотности распределения данных в окрестности объекта x_k. Чем ближе находится объект данных x_k к центру кластера, тем больше значение $F(x_k)$, так как больше объектов данных окружает x_k. На рис. 2 представлен график функции (2) для значений параметра $\gamma \in \{1, 5, 10, 20\}$.

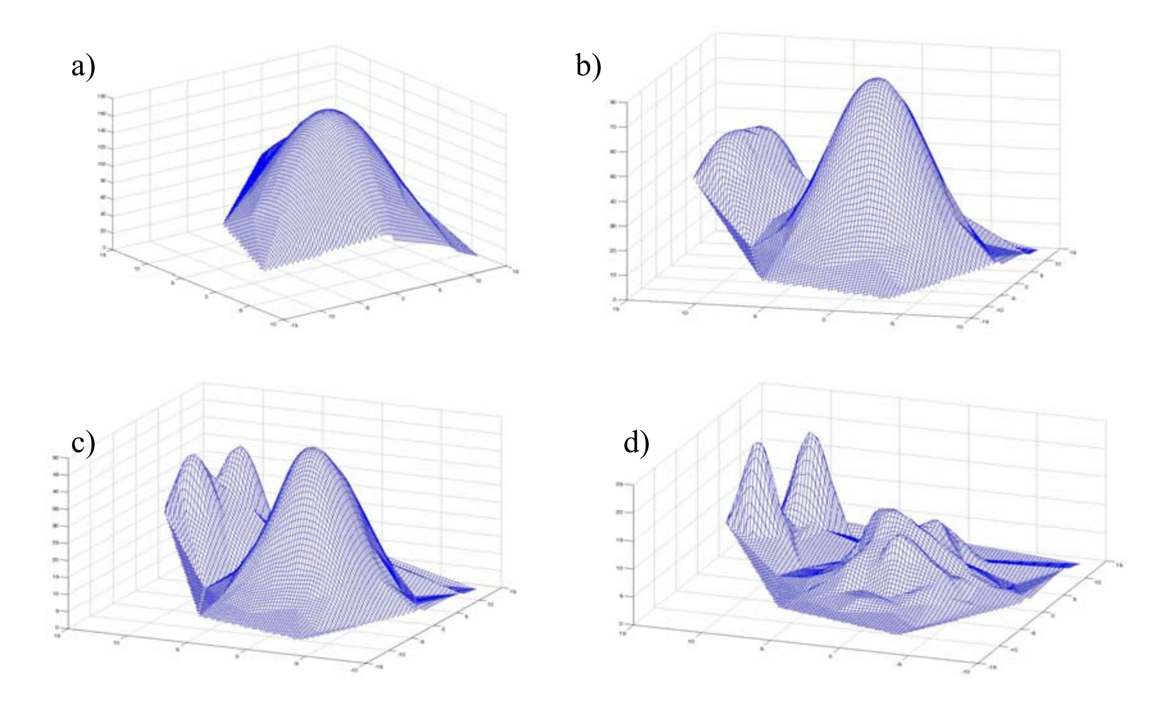


Рисунок 2 – Графики функции $F(x_k)$ для a) $\gamma = 1$; b) $\gamma = 5$; c) $\gamma = 10$; d) $\gamma = 20$

Согласно рис. 2 функция $F(x_k)$ имеет один локальный экстремум при $\gamma = 1$, при $\gamma = 5$ на графике функции $F(x_k)$ можно выделить два экстремума, и только при значении $\gamma = 10$ на графике появляются три локальных экстремума, соответствующие значениям функции $F(x_k)$ в центрах трех реально существующих кластеров набора

данных Data1. При дальнейшем возрастании параметра γ происходит выделение избыточных локальных экстремумов. Таким образом, параметр γ позволяет задавать границу кластера в окрестности точки x_k набора данных Data1. Проблема состоит в определении такого значения параметра γ, которое позволит определить действительную кластерную структуру данных для произвольного набора. Выбор слишком большого значения γ приведет к выделению большого количества малых кластеров в данных, где каждый объект будет представлять отдельный кластер, и функция $F(x_k)$ будет иметь большое количество локальных экстремумов. Слишком малое значение γ приведет к выделению единственного локального экстремума, даже в случае наличия в данных более одного кластера. При визуальном анализе рис. 2 значения функции $F(x_k)$ при $\gamma = 1$ и $\gamma = 5$ имеют значительные отличия, тогда как значения $F(x_k)$ при $\gamma = 5$ и $\gamma = 10$ отличаются незначительно. Поэтому для определения γ имеет смысл рассчитывать коэффициенты корреляции между значениями функции $F(x_k)$ для каждой пары последовательных значений параметра γ, где каждое последующее значение γ смещено относительно предыдущего на некоторый шаг (в нашем исследовании выбран шаг, равный 5). Как только значение из последовательности рассчитанных коэффициентов корреляции превысит некоторое пороговое значение, то можно считать, что функция $F(x_k)$ с соответствующим значением параметра γ определяет реальную кластерную структуру набора данных. Таким образом, можно считать, что последующее увеличение γ не привнесет полезной информации в оценку кластерной структуры данных, а только приведет к появлению дополнительных малых локальных экстремумов. На рис. 3 представлен график зависимости значений коэффициента корреляции ρ от значений параметра γ для набора данных Data1. Согласно результатам экспериментов в качестве порогового значения коэффициента корреляции целесообразно выбирать значение $\rho_П \in [0{,}97,\ 0{,}99]$.

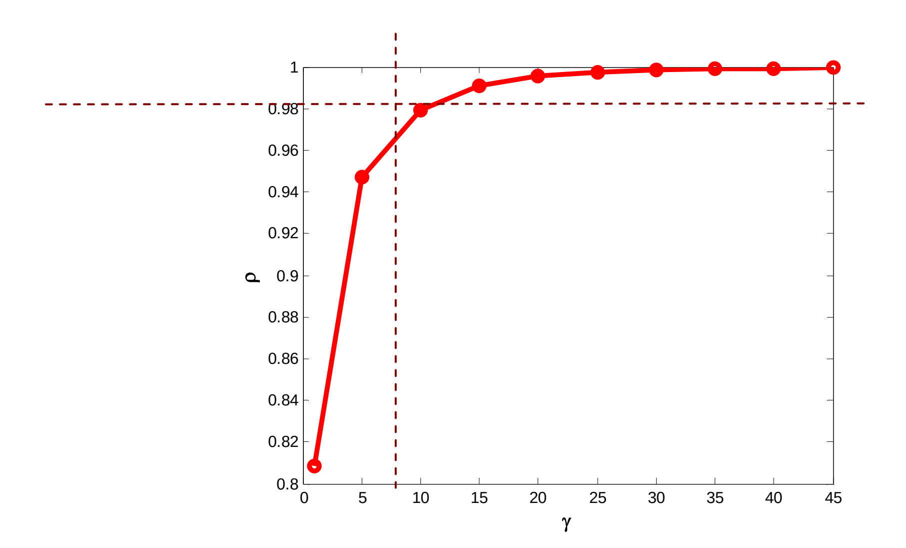


Рисунок 3 – График зависимости коэффициента корреляции ρ от параметра γ

Согласно рис. 3 для параметра γ выбирается значение 10, т.к. соответствующее значение ρ превышает пороговое значение $\rho_П \in [0{,}97,\ 0{,}98]$.

Кластеризация данных

Согласно предложенному авторами МКК после проведения оценки кластерной структуры анализируемого набора данных и выбора соответствующего значения параметра γ осуществляется кластеризация данных путем оптимизации функционала $F(v)$ (1). Необходимое условие максимизации функционала (1) следующее:

$$\frac{dF(v)}{dv_i}=0\,.$$

После дифференцирования функционала (1) относительно центров кластеров $v_i, i=1,\ldots,c$ получаем

$$\frac{dF(v)}{dv_i}=\sum_{j=1}^{n}2\cdot\frac{\gamma}{\delta}\cdot\left(x_j-v_i\right)\cdot\left(\exp-\frac{\left\|x_j-v_i\right\|^2}{\delta}\right)^{\gamma} \tag{3}$$

и следовательно необходимое условие максимума функционала (1) следующее:

$$v_i=\frac{\sum_{j=1}^{n}x_j\left(\exp-\frac{\left\|x_j-v_i\right\|^2}{\delta}\right)^{\gamma}}{\sum_{j=1}^{n}\left(\exp-\frac{\left\|x_j-v_i\right\|^2}{\delta}\right)^{\gamma}}\,. \tag{4}$$

Для поиска оптимальных значений $v_i, i=1,\ldots,c$ используется следующий итерационный алгоритм, состоящий из выполнения двух последовательных шагов, соответствующих выражению (4):

Инициализация начальных центров кластеров $v_i, i=1,\ldots,c$ *и задания значения* $e=0,01$ *для условия останова алгоритма.*

iter=0 – счетчик итерации алгоритма.

Шаг 1. Рассчитать значение меры сходства $f_{ij}^{iter+1}=\exp\left(-\frac{\left\|x_j-v_i^{iter}\right\|^2}{\delta}\right)$ *для каждого объекта данных* $x_j, j=1,\ldots,n$ *и каждого центра кластера* $v_i, i=1,\ldots,c$*;*

δ *– дисперсия всего набора данных.*

Шаг 2. Рассчитать новое значение центра кластера $v_i^{iter+1}=\frac{\sum_{j=1}^{n}f_{ij}^{iter+1}x_j}{\sum_{j=1}^{n}f_{ij}^{iter+1}}$.

Если $\max_i\left\|v_i^{iter+1}-v_i^{iter}\right\|<e$*, то окончание алгоритма и* $v_i^{кон}=v_i^{iter+1}, i=1,\ldots,c$

Иначе $v_i^{iter}=v_i^{iter+1}, i=1,\ldots,c$ *и перейти к шагу 1.*

Для того чтобы в процессе кластеризации были обнаружены все реально имеющиеся в наборе данных кластеры, в качестве начальных центров кластеров выбираются все объекты данных, т.е. полагается, что $c=n$ и $(v_1^0,v_2^0,\ldots,v_n^0)=\left(x_1,x_2,\ldots,x_n\right)$. Такая инициализация начальных центров кластеров гарантирует то, что в процессе

кластеризации будут найдены все локальные экстремумы функции $F(x_k)$ и количество найденных экстремумов будет соответствовать количеству реальных кластеров в наборе данных. Такого рода инициализация позволяет избежать произвольного определения количества кластеров в данных и является устойчивой к выбору начальных положений центров кластеров. Инициализация кластерного алгоритма с использованием большого числа кластеров применяется в методах прогрессивной кластеризации [18]. При проведении кластеризации авторами были исключены из рассмотрения в качестве начальных центров кластеров объекты данных, которые имели значение функции $F(x_k)$ меньше 10-го перцентиля значений $F(x_k), k = 1,\ldots,n$. Последовательное изменение положений центров кластеров в результате кластеризации набора данных Data1 приведено на рис. 4.

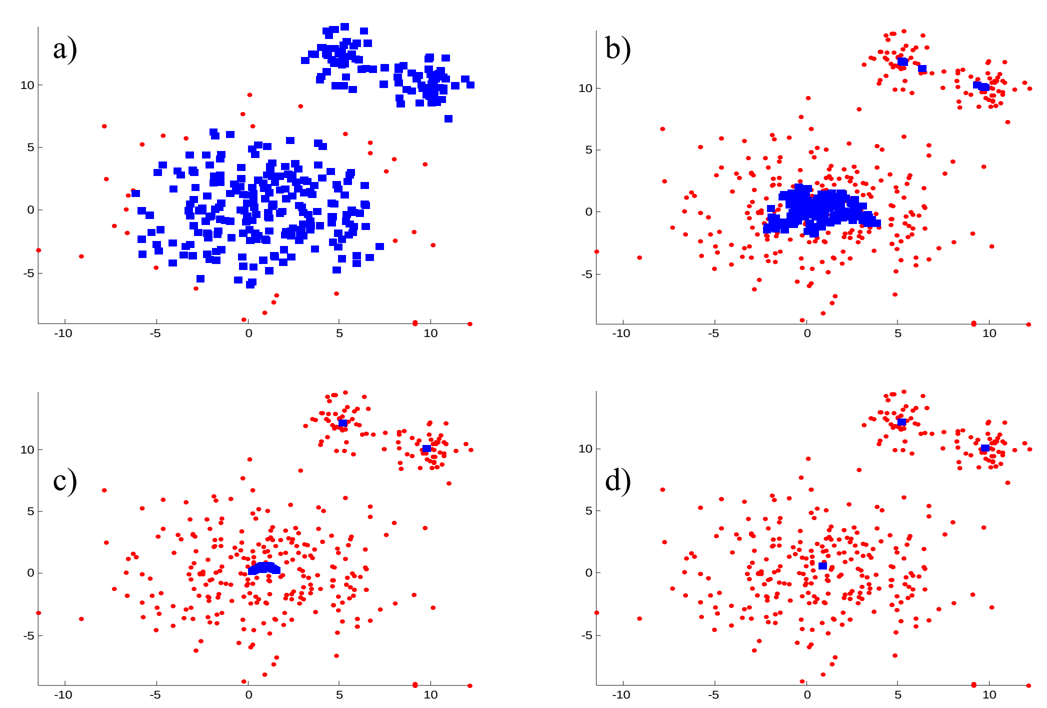


Рисунок 4 – Положение центров кластеров набора данных Data1 после а) 5 итераций; b) 10 итераций; c) 20 итераций; d) конечное положение

Как видно из рис. 4, оптимальное количество кластеров для набора данных Data1 равно трем. Таким образом, при выборе начальных положений центров кластеров $(v_1^0, v_2^0, \ldots, v_n^0) = (x_1, x_2, \ldots, x_n)$ алгоритм кластеризации является устойчивым к инициализации.

Для автоматизации определения оптимального количества кластеров c_{opt} конечные положения центров n кластеров группируются с использованием метода агломеративной иерархической кластеризации. Результат иерархической кластеризации конечных положений n центров кластеров для набора данных Data1 представлен на рис. 5, согласно которому можно выделить три хорошо различимых кластера данных.

Каждый кластерный центр $(v_1^0, v_2^0, \ldots, v_n^0) = (x_1, x_2, \ldots, x_n)$ соответствует отдельному объекту данных, таким образом иерархическая кластеризация n кластерных центров позволяет одновременно с оптимальным количеством кластеров c_{opt} определить их состав, т.е., какие объекты данных входят в состав каждого из полученных кластеров.

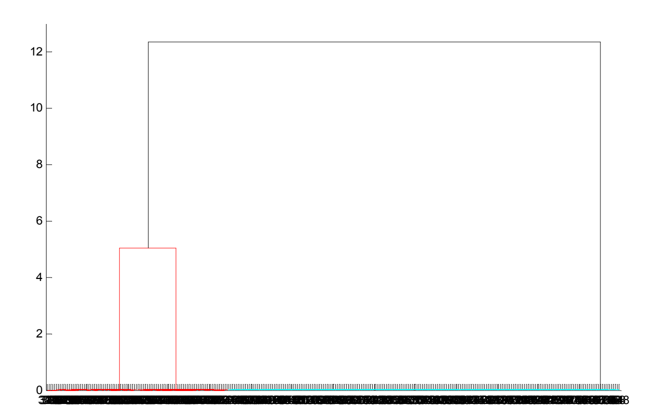


Рисунок 5 – Иерархическое дерево для набора данных Data1

Результаты кластеризации данных Data1 с использованием МКК приведены на рис. 6. Для сравнения на рис. 7 – 8 приведены результаты кластеризации с использованием метода FCM и метода PCM при $c = 3$ в случае выбора параметров границ кластеров $\eta_i, i = 1,\ldots,c$ согласно [14].

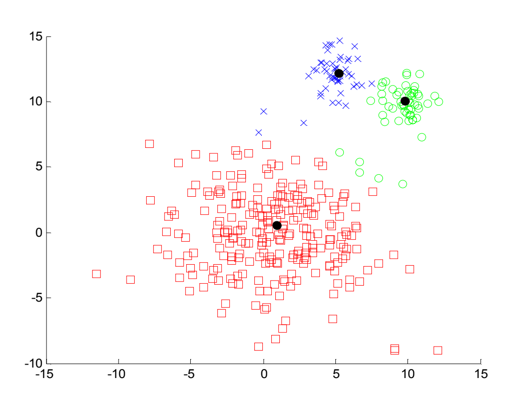


Рисунок 6 – Результат кластеризации набора данных Data1 с помощью МКК

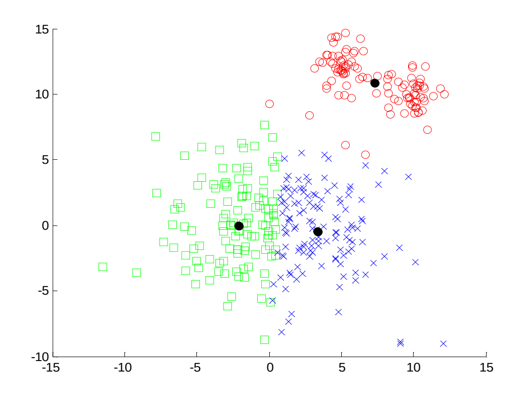


Рисунок 7 – Результаты кластеризации Data1 с помощью FCM

Из рис. 7 видно, что в результате кластеризации данных Data1 методом FCM наибольший по объему кластер разделяется на два кластера, а два малых кластера сливаются в один. Таким образом, метод FCM не позволяет распознать в данных различные по величине кластеры.

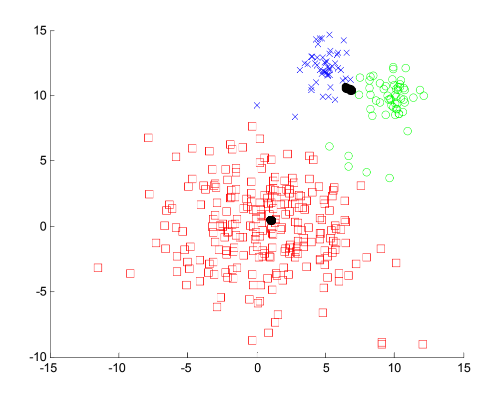


Рисунок 8 – Результаты кластеризации Data1 с помощью PCM

Из рис. 8 видно, что с использованием метода PCM набор данных Data1 разделяется на 3 кластера, однако центры меньших по объему кластеров смещены. Причиной такого результата является выбор параметров границ кластеров $\eta_i, i = 1,\ldots,c$.

Согласно предложенному авторами методу объекты данных самоорганизуются в процессе кластеризации, группируясь в областях с наибольшей плотностью данных, что позволяет достаточно точно определить положение центров кластеров и их количество.

Анализ устойчивости метода к выбросам в данных

В данном разделе проводится анализ устойчивости предложенного метода оценки кластерной структуры и кластеризации данных к выбросам в данных. В работе [17] было отмечено, что использование вместо евклидовой экспоненциальной меры расстояния между объектами позволяет повысить устойчивость кластерного алгоритма к выбросам в данных. Таким образом, предполагается, что предложенный авторами метод кластеризации, который основан на оптимизации функционала (1), являющегося суммой экспоненциальных расстояний между объектами данных и кластерными центрами обладает свойством устойчивости. Для проверки этой гипотезы нами был проведен эксперимент с использованием наборов данных Data2 и Data3, представленных на рис. 9.

Набор данных Data3 отличается от набора данных Data2 тем, что один из выбросов удален на большее расстояние от центров кластеров. Результаты кластеризации наборов данных с использованием предложенного метода приведены на рис. 10. Для сравнения результаты кластеризации наборов данных методами FCM и PCM приведены на рис. 11 – 12. Положения центров кластеров: реальные и полученные в результате кластеризации для наборов данных Data2 и Data3, приведены в табл. 1 – 2. Из рис. 10 видно, что положение центров кластеров достаточно устойчиво при удалении выброса от центров основных кластеров. В отличие от этого согласно рис. 11 при удалении выброса от реальных центров основных кластеров (набор Data3) один из центров, полученных после FCM кластеризации, расположен между основными кластерами, а второй – в точке выброса. Результаты PCM кластеризации также реагируют на местоположение выброса и оба центра основных кластеров сливаются в одну точку, расположенную между ними (рис. 12b). Согласно результатам экспериментов можно сделать вывод, что метод вероятностной кластеризации PCM в отличие от предложенного в работе метода не является устойчивым при удалении выброса от местоположения основных кластеров.

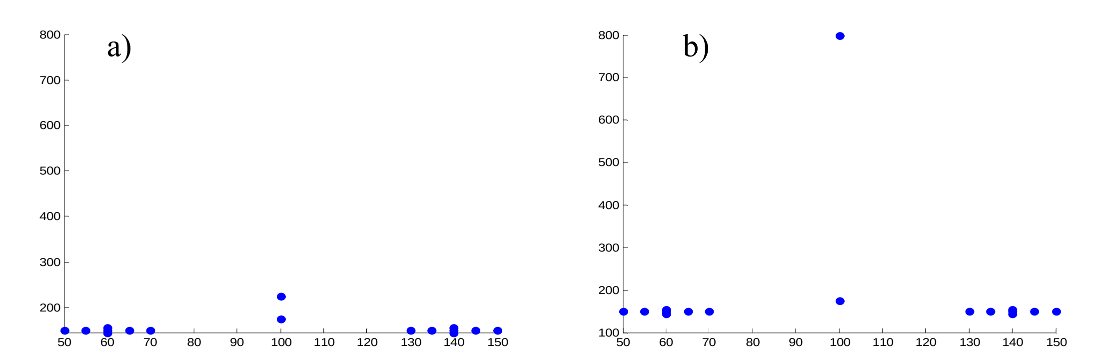


Рисунок 9 – Наборы данных, состоящие из двух кластеров и двух выбросов:
a) набор данных Data2; b) набор данных Data3

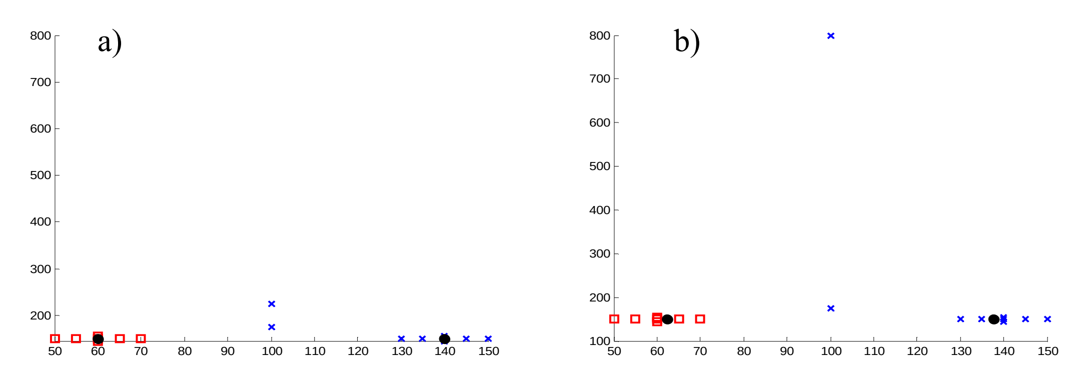


Рисунок 10 – Результат кластеризации с использованием МКК:
a) набор данных Data2; b) набор данных Data3

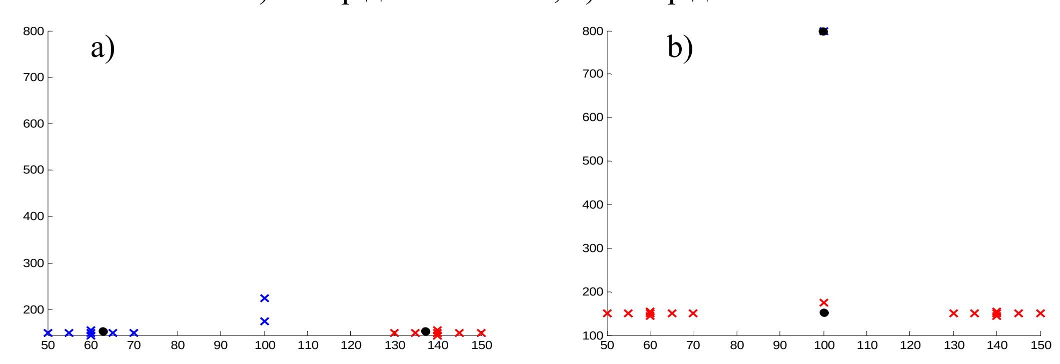


Рисунок 11 – Результат кластеризации с использованием FCM:
a) набор данных Data2; b) набор данных Data3

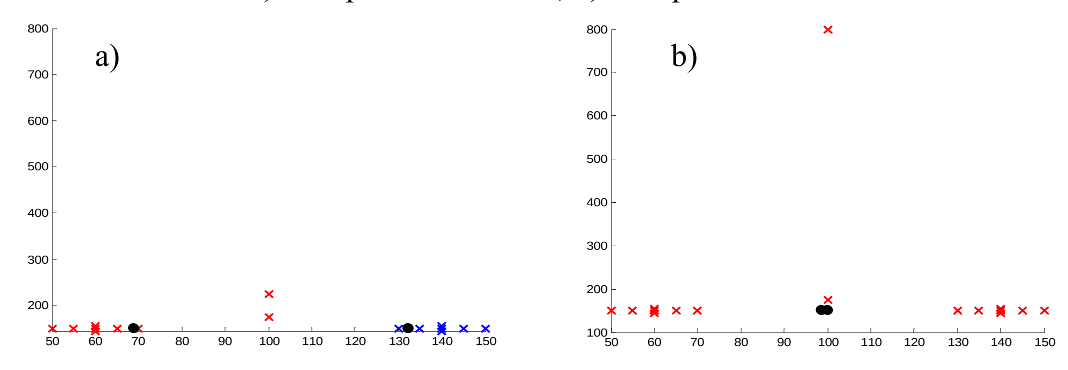


Рисунок 12 – Результат кластеризации с использованием PCM:
a) набор данных Data2; b) набор данных Data3

Таблица 1 – Координаты центров кластеров для набора данных Data2

№ клас-тера	Реальные координаты центров	Координаты центров, полученные МКК методом	Координаты центров, полученные методом FCM	Координаты центров, полученные методом PCM
1	(60, 150)	(60.015, 150.008)	(137.241, 153.408)	(132.072, 151.576)
2	(140, 150)	(139.984, 150.008)	(62.759, 153.409)	(69.347, 151.814)

Таблица 2 – Координаты центров кластеров для набора данных Data3

№ клас-тера	Реальные координаты центров	Координаты центров, полученные МКК методом	Координаты центров, полученные методом FCM	Координаты центров, полученные методом PCM
1	(60, 150)	(62.229, 150.766)	(100, 151.674)	(99.924, 152.367)
2	(140, 150)	(137.771, 150.766)	(100, 799.852)	(98.584, 152.616)

Заключение

Предложенный МКК метод оценки кластерной структуры и кластеризации данных является устойчивым к инициализации параметров кластеризации, к выбросам в данных и позволяет распознавать различные по объему кластеры. Кластерная структура и количество кластеров определяются в процессе самоорганизации объектов данных. При кластеризации осуществляется поиск таких значений центров кластеров $v_i, i = 1,\ldots,c$, которые максимизируют полную меру сходства объектов данных и кластерных центров. Для определения локально-оптимального количества кластеров и состава отдельных кластеров предлагается использовать агломеративную иерархическую кластеризацию значений центров кластеров, полученных в результате работы оптимизационного кластерного алгоритма. МКК метод может использоваться для осуществления предварительного анализа набора данных с целью выявления новых знаний или скрытых закономерностей, а также для последующего построения набора правил классификации, соответствующих отдельным кластерам. Направлением дальнейших исследований является разработка подхода к предварительной оценке кластерной структуры данных и выбора значений границ кластеров для реализации метода вероятностной кластеризации PCM.

Литература

1. Dubes R. Algorithms for Clustering Data / R. Dubes, A. Jain. – Prentice Hall, 1988.
2. Kaufman L. Finding Groups in Data: An Introduction to Cluster Analysis / L. Kaufman, P.J. Rousseeuw. – John Wiley and Sons, 1990.
3. Jain A.K. Data Clustering: A Review /A.K. Jain, M.N. Murty, P.J. Flynn // ACM Computing Surveys. – 1999. – Vol. 31, № 3. – P. 254-323.
4. McLachlan G.J. The EM Algorithm and Extensions / G.J. McLachlan, T. Krishnan. – John Wiley and Sons, 1997.
5. Fraley C. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis / C. Fraley, A.E. Raftery // The Computer J. – 1998. – Vol. 41, № 8. – P. 578-588.
6. Bezdek J.C. Pattern Recognition with Fuzzy Objectiv Function Algorithm / J.C. Bezdek. – Plenum Press, 1981.

7. Kohonen T. Learning Vector Quantization / T. Kohonen // Neural Network. – 1988. – Vol. 1. – P. 303.
8. Tsao E.C.K. Fuzzy Kohonen Clustering Net Works / E.C.K. Tsao, J.C. Bezdek, N.R. Pal // Pattern Recognition. – 1994. – Vol. 27. – P. 757-764.
9. Hartuv E. A Clustering Algorithm Based on Graph Connectivity / E. Hartuv, R. Shamir // Information Processing Letters. – 2000. – Vol. 76, № 4-6. – P. 175-181.
10. Jiang D. DHC: A Density-Based Hierarchical Clustering Method for Time-Series Gene Expression Data / D. Jiang, J. Pei, A. Zhang // Proc. BIBE2003: Third IEEE Int'l Symp. Bioinformatics and Bioeng. – 2003.
11. Ertoez L. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data / L. Ertoez, M. Steinbach, V. Kumar // Proceedings of the SIAM International Conference on Data Mining. – 2003.
12. McQueen J.B. Some Methods for Classification and Analysis of Multivariate Observations / J.B. McQueen // Proc. Fifth Berkeley Symp. Math. Statistics and Probability. – 1967. – Vol. 1. – P. 281-297.
13. Fuzzy Cluster Analysis: Methods for Classification Data Analysis and Image Recognition / Hoppner F., Klawonn F., Kruse R., Runkler T. – New York : Wiley, 1999.
14. Krishnapuram R. A Possibilistic Approach to Clustering / R. Krishnapuram, J.M. Keller // IEEE Trans. Fuzzy Systems. – 1993. – Vol. 1. – P. 98-110.
15. Dave R.N. Characterization and Detection of Noise in Clustering / R.N. Dave // Pattern Recognition Letters. – 1991. – Vol. 12. – P. 657-664.
16. Жук Е.Е. Устойчивость в кластер-анализе многомерных данных / Е.Е. Жук, Ю.С. Харин. – Мн. : Белгосуниверситет, 1998. – 240 с.
17. Wu K.L. Alternative c-means clustering algorithms / K.L. Wu, M.S. Yang // Pattern Recognition. – 2002. – Vol. 35. – P. 2267-2278.
18. Frigui H. A Robust Competitive Clustering Algorithm with Applications in Computer Vision / H. Frigui, R. Krishnapuram // IEEE Trans. Pattern Analysis and Machine Intelligence. – 1999. – Vol. 21. – P. 450-465.

Н.А. Новосьолова, І.Е.Том

Метод оцінки кластерної структури і кластеризації даних

У статті розглядається проблема розробки методів кластеризації, які є стійкими до ініціалізації (кількість кластерів і початкові параметри кластерів), до різних за об'ємом кластерів, до викидів в даних. Пропонується метод оцінки кластерної структури і кластеризації даних, який заснований на розрахунку значень близькості об'єктів даних в багатовимірному ознаковому просторі. Метод є стійким до ініціалізації параметрів кластеризації, до викидів в даних і дозволяє визначати кластерну структуру і кількість кластерів в ході самоорганізації об'єктів даних.

N.A. Novoselova, I.E. Tom

Method of Evaluation of Clustering Structure and Data Clustering

The paper is devoted to the problem of development of the clustering methods, which are robust to initialization (number of clusters and initial cluster parameters), to the different cluster volumes, to the outliers. It is proposed a method for estimation of cluster structure and clustering of data, based on the evaluation of similarity measure between data objects in multidimensional space. The proposed method is robust to initialization of clustering parameters, to outliers and allows definition of cluster structure and number of clusters in the data self-organizing process.

Статья поступила в редакцию 21.07.2010.